Checklist for writing a thesis or paper

Anne-Laure Boulesteix

November 21st 2013

Formulas and mathematical notations

- Formulas should be integrated in the text, i.e. considered as part of a sentence and not isolated from the text.
- When the formula is the end of the sentence, it should end with a dot.
- Otherwise, it often makes sense to put a comma at the end of the formula (but not always: just read the sentence for yourself to decide which punctuation is better).
- All notations involved in the formula should be defined somewhere, at latest in the sentence following the formula, for example:
  We obtain

  XXXX = YYYY,

  where XXXX denotes … .

- If a mathematical notation is used only once in the manuscript, consider removing it. Mathematical notations help a lot in many cases, but they may also unnecessarily generate confusion (keep in mind that it also represents an effort for the reader to remember all the introduced notations).
- Do not use the same notation for more than one concept in the whole manuscript.
- When defining mathematical notations, stick to generally accepted conventions whenever possible except if you have really good reasons not to do so.
- A useful convention is to write vectors and matrices with bold characters. It is not mandatory to follow this convention, but it is mandatory to keep consistent (i.e. write either all vectors and matrices bold or none of them).
- Bold greek letters can be obtained in latex with the command \bm{} from the package 'bm'.
- Equations can be referred to in several ways, for example "in Equation (1)", "in Eq. (1)", "in (1)". Choose one option and stick to it.

Bibliography

- All papers and books appearing in the bibliography section should be referred to somewhere in the main manuscript.
- Items referring to articles should include information on authors, year of publication, title, journal, volume, first page, last page.
- The only exception to the above rule is for journals (mostly online journals) whose articles are numbered by a single number and not by page numbers, for example BMC journals.

- The issue number can also be given in parentheses after the volume number, but it is not mandatory. If you decide to include it, include it for *all* papers.
- Items referring to books should include information on authors, year of publication (be careful: some books are edited several times), title, publisher (for example: Springer), place of publication (for example: New York).
- Avoid referring to websites whenever possible, i.e. try to find another source (journal, book, etc) where the information from the website are also available. If it is not possible, give the complete URL of the website and when it was accessed for the last time.
- For unpublished works and PhD theses, give as much information as possible (number of the technical report, university where it was published, etc) as well as the URL.
- Googlescholar provides the bibtex code of almost all scientific publications. You can use it to save time, but do not forget to check it carefully: it often includes small formatting errors.
- Words that are part of journal names should begin with a capital letter (this is a common error in googlescholar, see above topic).
- If you decide to abbreviate the journals' names, do it consistently for all journals and use the official abbreviations.

Citations

Correct ways to refer to literature:

- In Smith et al (2013) …
- … by Smith et al (2013)
- Smith et al (2013) suggest …
- … has been suggested previously (Smith et al, 2013).
- … has been suggested previously [1].
- In Smith et al [1] …
- … by Smith et al [1]
- Smith et al [1] suggest …
- … (see Smith et al, 2013) -> this can be obtained with the command \citep[see][]{smith2013}

Incorrect ways to refer to literature:

- In (Smith et al, 2013)
- … by (Smith et al, 2013)
- … has been suggested previously (Smith et al (2013))
- In [1]…
- … by [1]
- [1] suggest …
- … (see Smith et al (2013))

Style

- Do not write too long sentences, especially if you are a beginner in writing. Do not hesitate to split sentences following the rule "one idea one sentence".

- Avoid repetitions of words in the same sentence or in consecutive sentences. This rule applies only to words that do not refer to precise scientific concepts.
- Words referring to precise scientific concepts should be repeated as often as needed and not replaced by other variants. It would make the manuscript more confusing! Example: if you use the term “predictor variable”, use it throughout the manuscript and do not replace it by “variable”, “predictor”, “explicative variable”, “independent variable” and so on. The same holds, e.g., for “prediction method”, which should not be replaced arbitrarily by “prediction algorithm”, “learning method”, etc. Stick to your terminology.
- Avoid footnotes whenever possible (many journals do not accept them).
- Use the different tenses consistently. The tense might change throughout the manuscript, for example present tense in the methods section and past tense in the results section. But one should stick to the chosen scheme. In statistical literature, the usual convention is to write in present tense throughout the manuscript (whereas medical papers typically use past tense for the results section).
- When referring to tables, figures, sections, subsections, use capitals for the first letter (“Figure”, “Table”, “Section”, “Subsection”) if (and only if) it is followed by a number. Example: “The method is described in Section 2” but “The next section presents the method”.

Structuring

- If there is a subsection 2.1 (or subsubsection 2.1.1), there should also be a subsection 2.2 (or subsubsection 2.1.2). Otherwise, do not divide into (sub)subsections.
- Ideally, the subsection titles should have approximately the same level of detail. For example, a subsection title consisting of one single word followed by a subsection title that does not fit in one line would look odd (although there is nothing really wrong in doing so).
- Write transitions between the (sub)sections.
- Defer non-crucial information to the appendix.
- But do not forget to refer to these information in the main text.
- The number of subsections in a section should not exceed, say, 6 ou 7.

Proofreading the manuscript

- Run a spellchecker just before printing/submitting your work. If you run the spellchecker too early, new errors are likely to be introduced in the meantime.
- Check for double words, for example “in the the paper”: most spellcheckers will not find such errors.
- Consider using a grammar checker, for example by copying your latex code into Word (the grammar spellchecker would find double words and other problems that are not found by a standard spellchecker).
- When proofreading the manuscript, do not forget parts like figure and table legends, bibliography. Also these parts should be correct.
- Repeat proofreading on several days whenever possible. Mistakes are most likely to be discovered by a fresh mind.

- Many people find that it is easier to find mistakes in a printed document than on a screen.

Figures/Tables

- The legend of tables and figures should include all important information: whether real or simulated data, which setting/dataset, which method, which variant, etc. Someone familiar with the thesis/paper should be able to understand the content of the table or figure without reading the text again.
- In particular: make sure that all tables and figures have different legends. The differences between related tables and figures should be recognized by reading the legend.
- The difference between panels of a figure should be recognized at first glance without reading the text and if possible without even reading the legend.
- The legend might also include a sentence summarizing the main result displayed in the figure.
- Axes should be adequately labeled, even if you think that "it should be obvious to the reader" (it is often not that obvious).
- Check that the axis labels and legends are large enough to be read when printed out.
- Different curves should be different enough to be distinguished from each other without much effort.
- In particular, it is recommended to use line styles/colors that can be distinguished when printed with a black&white printer, since most readers will not print the manuscript in colors.
- Figures and tables should all be referred to in the text. Examples of correct formulations:
    - Figure 1 displays the boxplot of…
    - The results of simulation XX are shown in Figure 1.
    - As can be seen from Figure 1, …
    - It can be observed from Figure 1 that …
    - When looking at Figure 1 it is striking/obvious that…
    - Figure 1 illustrates the results of simulation XXX

Formatting

- Avoid too short or too long paragraphs. Possible rule of thumb: paragraphs should have at least 4-5 lines and there should be at least 3 paragraphs in a page.
- In the very final version of a thesis, check that you do not have section titles/single lines at the end of pages or single lines at the top of pages.

Explaining what you do

- Clearly state in the introduction what the contributions of your thesis/paper are.
- It means that you should very clearly explain what is already existing and what your manuscript adds to the state-of-the art.
- Explain why your contribution is needed and important.
- In the methods section, clearly state whether you are reviewing existing literature or presenting your own new concepts – even if you think it is obvious! Do not forget that your work might also be read by laypersons who are not very familiar with the "existing" methods.

- Ideally, the review of existing methods and the new contributions should be presented in different subsections whose titles should reflect the review or new character, respectively.

English language issues

- For german native speakers: In English the use of commas before subordinate clauses is not systematic, for example we would write “the method that is described in this article” and not “the method, that is described in this article”.
- For german native speakers: in English, adjectives and adverbs usually do not have the same form, for example typical (adjective) vs. typically (adverb). Since this differentiation does not exist for most adjectives and adverbs in German, check adjectives and adverbs with particular attention with respect to this problem.
- Check that verbs following subjects in singular form end with “s” and that verbs following subjects in plural form do not end with “s”.
- Words like “rather”/“quite”, “pretty” are considered as slang. You can use them in oral presentations, but should avoid them as far as possible in a manuscript (“rather … than …” is ok: what is not ok is to use “rather” in the sense of “quite”)
- “can not” is written as “cannot”.
- Abbreviations like “can’t”, “isn’t”, etc should be avoided in manuscripts.
- The word “data” can either be considered as plural (data are) or singular (data is). Choose the convention you prefer but stick to it throughout the manuscript.
- The word “dataset” can also be written as “data set”. Choose the convention you prefer but stick to it throughout the manuscript.